Procesamiento del lenguaje Natural, nº 41 (2008), pp. 301-302 recibido 7-05-2008; aceptado 16-06-2008

Herramientas de anotación de corpus de habla espontánea del Laboratorio de Lingüística Informática de la UAM

Toolbox for annotating spontaneous speech corpora (Computational Linguistics Lab – UAM)

Antonio Moreno Sandoval
Laboratorio de Lingüística Informática
UAM
Antonio.msandoval@uam.es

José Ma. Guirao Miras
Dept. de Lenguajes y Sistemas Informáticos
UGranada
jmguirao@ugr.es

Doroteo Torre Toledano
Dept. Ingeniería Informática
UAM
Doroteo.torre@uam.es

Resumen: Presentamos un sistema de anotación fonológica, silábica y morfosintáctica (incluyendo categoría sintáctica, lema y rasgos morfológicos) especialmente adaptado para corpus orales. Todas las herramientas se han desarrollado y validado en corpus de habla espontánea (C-ORAL-ROM, CHIEDE, CORLEC).
Palabras clave: anotación, fonología, sílaba, lematización.

Abstract: We show a toolbox for linguistic annotation (including phonology, sillabification, part of speech, lemma and morphological features) especially adapted to Spanish spoken corpora. These tools have been developed and validated against several spontaneous speech corpora compiled by the Laboratorio de Lingüística Informática-UAM: C-ORAL-ROM, CHIEDE, CORLEC
Keywords: Corpus annotation, phonology, syllable, lemmatization, PoS tagging.

1 Características del sistema

1.1 La transcripción de corpus orales

Los corpus orales son mucho más complejos de compilar que los corpus escritos y en especial exigen una dedicación intensiva de los transcriptores: estimamos que para cada hora de grabación se necesitan unas 40 horas de trabajo especializado. Las tareas incluyen:

1) preparación del entorno de grabación,
2) registro de las conversaciones o monólogos,
3) tratamiento digital de las grabaciones hasta que se obtiene el fichero fuente de sonido,
4) transcripción manual por lingüistas especializados,
5) anotación prosódica manual (pausas, disfluencias, solapamientos, etc.),
6) revisión de la transcripción y anotación prosódica por un lingüista distinto,
7) revisión conjunta de los dos lingüistas en las cuestiones con discrepancia,
8) alineamiento manual de cada segmento o utterance, es decir, del segmento sonoro con su transcripción.

Una vez terminado el proceso de transcripción (ortográfica y prosódica) se comienza con la anotación de la información propiamente lingüística. Digamos que este proceso preliminar correspondería a la compilación de un corpus escrito con la introducción de metadatos en la cabecera de cada texto.

1.2. La anotación del nivel fonológico

Los corpus orales exigen este nivel de anotación, a diferencia de los escritos. Estos corpus se suelen emplear para dos tipos de tareas básicas:

1. entrenamiento de sistemas de reconocimiento de habla
2. base de datos para descripción de las características de la lengua oral

ISSN: 1135-5948

En el primer caso, los corpus de habla espontánea sirven de base de datos acústica y de modelo de lengua. Para ello tienen que estar en formato "fonológico": cada fonema lleva un símbolo que lo identifique inequívocamente. La transcripción ortográfica es inservible.

En principio sería posible realizar la transcripción fonológica manualmente, pero requeriría tanto esfuerzo en términos de tiempo que no sería viable para un corpus de más de 50.000 palabras[1].

Por ello, empleamos un transcriptor fonológico que transcribe automáticamente cualquier texto escrito en ortografía castellana estándar. El transcriptor se ha descrito en varias publicaciones.

La tasa de error estimada es menor al 2 por ciento y se concentra exclusivamente en las palabras de ortografía no castellana (es decir, extranjerismos, como "web"), nombres propios no castellanos ("John") y acrónimos ("SEPLN"). La manera de resolver estos problemas es mediante la inclusión en una lista de excepciones, pero obviamente esa lista es muy incompleta.

El transcriptor además de traducir el texto a formato "fonológico" hace una silabificación y asignación de acento fonológico. No se ha realizado una evaluación exhaustiva pero sabemos que funciona con un nivel de precisión similar al fonológico, siempre que se tome cada palabra aisladamente. La juntura externa entre palabras no se trata de momento.

1.3. La anotación morfosintáctica

Como es bien sabido, esta anotación es el paso inicial esencial en el procesamiento automático, puesto que proporciona la información de entrada al nivel sintáctico (categoría y rasgos de concordancia) y al nivel semántico (el lema).

La anotación morfosintáctica del habla espontánea comparte los mismos problemas y requisitos que en los sistemas para textos escritos:

- reconocimiento de multi-words
- desambiguación de varios análisis posibles
- tratamiento de palabras nuevas o desconocidas

Partimos de un tagger diseñado y entrenado para textos escritos (Grampal, 1991) y le hemos añadido algunas especificaciones nuevas para adaptarlo a los textos orales:

- un tokenizador especial para los textos transcriptos, incluyendo reconocimiento de disfluencias (alargamientos vocálicos, interrupciones de palabras, sonidos paralingüísticos...),
- entrenamiento para desambiguación de análisis dentro de un contexto oral,
- inclusión de la categoría Marcador Discursivo ("es decir", "o sea", "bueno", "vale") necesaria para la sintaxis oral,
- añadido de un módulo para el tratamiento de diminutivos ("cafetito") y neologismos ("megahortera"), mucho más frecuentes en la lengua oral que en la escrita.

2 *La demostración*

El sistema de demostración permitirá a los asistentes comprobar el funcionamiento de las herramientas mediante una conexión remota a nuestro servidor.

Hay dos modos de consulta:

a. introducción de palabras aisladas
b. introducción de un texto

En el primer caso, el demostrador muestra el resultado en una presentación "gráfica". Cuando se introduce un texto, el demostrador proporciona una salida etiquetada en xml. De esta manera se muestran distintas posibilidades.

Bibliografía

Cresti, E., y M. Moneglia (eds.). 2005. *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam, John Benjamins.

Moreno, A. y J.M. Guirao. 2006. Morpho-syntactic tagging of the Spanish C-ORAL-ROM Corpus. In *Spoken Language Corpus and Linguistic Informatics*. Amsterdam, John Benjamins.

[1] Si estimamos una media de 5 fonemas por palabra, nos saldría un mínimo de 250.000 tokens-fonemas.